Published in final edited form as:

Genet Epidemiol. 2009 September ; 33(6): 463–478. doi:10.1002/gepi.20399.

Phase uncertainty in case-control association studies

Sungho Won, Sulgi Kim, and Robert C. Elston

Department of Epidemiology and Biostatistics, Case Western Reserve University

Abstract

The possible evidence for association comprises three types of information: differences between cases and controls in allele frequencies, in parameters for Hardy Weinberg disequilibrium (HWD), and in parameters for linkage disequilibrium (LD). LD between marker and disease alleles results in a difference in at least one of the three types of parameters [Won and Elston, 2008]. However, the parameters for LD require knowledge about phase, which is usually unknown, making the LD contrast test without modification infeasible in practice. Methods for handling phase uncertainty are: (1) the most probable haplotype pair for each individual can be considered as the true phase; (2) a weighted average of haplotypes can be used; (3) we can consider the composite LD, which does not require any information about phase. We compare these methods to handle phase uncertainty in terms of validity and efficiency, and the effect on them of HWD in the population, at the same time confirming results for the three types of information. When the LD between markers is high, the LD contrast test that uses a weighted average of haplotypes or the most probable haplotypes to calculate the LD is recommended, but otherwise the LD contrast test that uses the composite LD is recommended. We conclude that, even though the difference in allele frequencies is usually the most informative test except in the case of a recessive disease, the LD contrast test can be more powerful if the markers are dense enough.

Keywords

linkage disequilibrium; haplotype phase; self replication

Introduction

The goal of population association studies is to identify genetic polymorphisms that distinguish among individuals with different disease status, and it has been shown that an association between marker and disease alleles can be detected as a result of gametic phase disequilibrium for linked loci, which is what we shall mean here by linkage disequilibrium (LD). Thus, today LD can play a key role in genetic epidemiology and the advent of SNPs has enabled association studies at the genome-wide level. However, notwithstanding their efficiency, the huge number of SNPs has generated the problem of multiple testing and statistical methods to combat this issue have been investigated. SNP reduction, sometimes called self-replication [Van Steen et al., 2005; Zheng et al., 2007] is a possible solution to this problem when we have two independent statistics that can be used to test the same biological hypothesis. For example, suppose we have 500,000 equally-spaced SNPs in a genome-wide association (GWA) study, i.e. SNPs about 6Kb apart, then correction for multiple tests would require a Bonferroni-corrected p-value of $0.05/500,000$ for genome-wide significance at the 0.05 level, which is difficult to meet in practice. However, suppose we have two independent statistics (t_1 and t_2) and their p-values are u_1 and u_2 . Then, if 1%

of the SNPs are selected by screening based on either of the two independent statistics, and only the selected SNPs are used for a testing phase with the other statistic, we now only need a Bonferroni-corrected p-value of 0.05/5,000 at this testing phase to have genome-wide significance at the 0.05 level. This self-replication can be reformulated in terms of combining p-values. The rejection region for t_1 (or t_2), in this example of self-replication is asymptotically $u_1 < 0.05/5,000$ with $u_2 < 0.01$ (or $u_2 < 0.05/5,000$ with $u_1 < 0.01$). We can also use Fisher's method or Liptak's method [Liptak, 1958] for combining p-values (see Figure 1 for their rejection regions). As a result, better approaches than self-replication can be attained as long as we use the optimal method for combining u_1 and u_2 ; and we have shown that the most powerful method can be obtained under certain circumstances even though no uniformly most powerful test is possible [Won et al., 2008].

It has been shown that LD between marker and disease alleles can result in differences of the following parameters between cases and controls: allele frequencies, the parameters for Hardy-Weinberg disequilibrium (HWD) and the parameters for LD between markers [Won and Elston, 2008]. The fact that the phase-known genotype frequencies can be parameterized in terms of these three parameters shows they can each be used for association studies. In principle, there is evidence for association if any one of the parameters is different between cases and controls provided certain conditions - such as no genotyping error - hold, and the three types of information can have different levels of power according to the disease mode of inheritance and LD structure. However, even though these results provide intuition about an optimal strategy for analysis, their application to a practical situation is restricted because the usual LD contrast test requires information about phase, which is often unknown. Three ways of overcoming the phase uncertainty have been considered. First, the most probable haplotype pair for each individual can be considered as the true phase. Second, a weighted average of haplotypes can be used. Third, we can use the composite LD [Wang et al., 2007; Zaykin, 2004; Zaykin et al., 2006], which does not require any information about phase. The validity and efficiency of these three approaches have not yet been investigated. We therefore compare these approaches to handle phase uncertainty in terms of their type I error and power according to disease mode of inheritance and SNP density. At the same time, we investigate the effect of HWD in the population on the analysis. Our results show that, whereas the composite LD always preserves the type I error with reduced power, the methods that use the most probable haplotype or weighted haplotypes for each individual can have better power with type I error preserved provided dense markers are used. We further show that, when the markers are dense, the LD contrast test can be more powerful than a test based on allele frequency differences.

Methods

Notation

Let X_{ij} and Y_{ij} be random variables for marker loci A and B defined by

$$X_{ij} = \begin{cases} 1 & \text{if an allele at locus } A \text{ in haplotype } j \text{ of individual } i \text{ is } A_1 \\ 0 & \text{if an allele at locus } A \text{ in haplotype } j \text{ of individual } i \text{ is } A_2 \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{if an allele at locus } B \text{ in haplotype } j \text{ of individual } i \text{ is } B_1 \\ 0 & \text{if an allele at locus } B \text{ in haplotype } j \text{ of individual } i \text{ is } B_2 \end{cases}$$

where $j = 1$ (2) indicates a maternal (paternal) haplotype. For given marker loci A and B , let

$p_{A_k}, p_{A_k A_{k'}}, p_{A_k B_k}, p_{A_k B_k A_{k'} B_{k'}}, p_{A_k B_k}^{A_{k'} B_{k'}}$ and $p_{A_k \cdot B_{k'}}$ be respectively the frequencies of allele A_k , genotype $A_k A_{k'}$, haplotype $A_k B_k$, phase-unknown genotype $A_k B_k A_{k'} B_{k'}$, phase-known genotype $A_k B_k / A_{k'} B_{k'}$, and the joint frequencies of alleles A_k and $B_{k'}$ of two different gametes

in the population, where, for the SNPs that are used, there are usually only two alleles, A_1 (B_1) and A_2 (B_2). Also, for locus A , let $p_{A_k|case}$ and $p_{A_k|control}$ be the frequencies of allele A_k in the case group and control group, respectively, and let $p_{A_k A_{k'}|case}$ ($p_{A_k A_{k'}|control}$) and $p_{A_k B_{k'}|case}$ ($p_{A_k B_{k'}|control}$) be the analogous frequencies of the genotypes $A_k A_{k'}$ and joint frequencies of alleles A_k and $B_{k'}$ in two different gametes. As measures of HWD, let $d_A \equiv p_{A_1 A_1} - (p_{A_1})^2$, $d_{A|case} \equiv p_{A_1 A_1|case} - (p_{A_1|case})^2$ and $d_{A|control} \equiv p_{A_1 A_1|control} - (p_{A_1|control})^2$ for the population, cases and controls, respectively. For haplotype level HWD, we define the following:

$$\begin{aligned} d_{12}^{AB} &\equiv p_{A_1 B_1} p_{A_1 B_2} - \frac{1}{2} p_{A_1 B_1}^2, & d_{13}^{AB} &\equiv p_{A_1 B_1} p_{A_2 B_1} - \frac{1}{2} p_{A_2 B_1}^2 \\ d_{14}^{AB} &\equiv p_{A_1 B_1} p_{A_2 B_2} - \frac{1}{2} p_{A_2 B_2}^2, & d_{23}^{AB} &\equiv p_{A_1 B_2} p_{A_2 B_1} - \frac{1}{2} p_{A_1 B_2}^2 \\ d_{24}^{AB} &\equiv p_{A_1 B_2} p_{A_2 B_2} - \frac{1}{2} p_{A_2 B_2}^2, & d_{34}^{AB} &\equiv p_{A_2 B_1} p_{A_2 B_2} - \frac{1}{2} p_{A_2 B_1}^2. \end{aligned}$$

Then $d_A = -(d_{13}^{AB} + d_{14}^{AB} + d_{23}^{AB} + d_{24}^{AB})$ and $p_{A_1 B_1} p_{A_1 B_1} - p_{A_1 B_1}^2 = -d_{12}^{AB} - d_{13}^{AB} - d_{14}^{AB}$.

To parameterize LD, let $\Delta_{A_1 B_1} \equiv p_{A_1 B_1} - p_{A_1} p_{B_1}$, $\Delta_{A_1 B_1|case} \equiv p_{A_1 B_1|case} - p_{A_1|case} p_{B_1|case}$ and $\Delta_{A_1 B_1|control} \equiv p_{A_1 B_1|control} - p_{A_1|control} p_{B_1|control}$ be the corresponding parameters between the two loci. The composite LD parameters, which do not require any information about phase, are respectively defined as $\Delta_{A_1 B_1}^C = p_{A_1 B_1} + p_{A_1 B_1} - 2 p_{A_1} p_{B_1}$, $\Delta_{A_1 B_1|case}^C = p_{A_1 B_1|case} + p_{A_1 B_1|case} - 2 p_{A_1|case} p_{B_1|case}$ and $\Delta_{A_1 B_1|control}^C \equiv p_{A_1 B_1|control} + p_{A_1 B_1|control} - 2 p_{A_1|control} p_{B_1|control}$ in the population, cases and controls, respectively. Finally, it should be kept in mind that this notation is also applied to a disease locus (D) by correspondingly replacing the notation for the locus and alleles.

Phase uncertainty

When there are markers around a disease locus D , where D_1 (D_2) denotes a disease (normal) allele, association can be detected from the LD between the marker and the disease alleles, and we have shown that the information for association in case-control studies consists of three different parts [Won and Elston, 2008]:

I_a : trend (or difference) in allele frequencies

I_{HWD} : trend (or difference) in allele frequencies

I_{LD} : trend (or difference) in LD.

For each type of information, we define the three principal statistics as follows [Nielsen et al., 1998; Nielsen et al., 2004; Sasieni, 1997; Song and Elston, 2006]:

$$\begin{aligned} S_a &= \frac{\hat{p}_{A_1|case} - \hat{p}_{A_1|control}}{\sqrt{\text{var}(\hat{p}_{A_1|case} - \hat{p}_{A_1|control})}} S_{HWD} = \frac{\hat{d}_{A|case} - \hat{d}_{A|control}}{\sqrt{\text{var}(\hat{d}_{A|case} - \hat{d}_{A|control})}} \\ S_{LD} &= \frac{\hat{\Delta}_{A_1 B_1|case} - \hat{\Delta}_{A_1 B_1|control}}{\sqrt{\text{var}(\hat{\Delta}_{A_1 B_1|case} - \hat{\Delta}_{A_1 B_1|control})}}. \end{aligned}$$

However, haplotypes are usually unknown and only the genotype at each locus is observed. Three possible ways that handle phase uncertainty have been used. First, the most probable haplotype pair for each individual can be estimated and considered as the true phase. Second, a weighted average of haplotypes can be used. Third, we can use the composite LD

[Wang et al., 2007; Zaykin, 2004; Zaykin et al., 2006], which does not require any information about phase. We denote these methods as follows:

I_{LD}^m : trend (or difference) in LD using the most probable haplotypes

I_{LD}^w : trend (or difference) in LD using weighted haplotypes

I_{LD}^c : trend (or difference) in composite LD.

Also, as has been known for some time, if the variance for I_{LD} is used in place of the true variances of the three corresponding statistics, the type I error can be either inflated or deflated; this will also be seen in our results. Thus, the variances of the statistics corresponding to I_{LD}^m , I_{LD}^w and I_{LD}^c need to be derived.

First, we consider the most probable haplotypes and weighted averages of haplotypes obtained via haplotype frequency estimation. If the haplotype frequencies are known, then, letting $w = p_{A_1B_1}p_{A_1B_1}/(p_{A_1B_1}p_{A_1B_1} + p_{A_1B_2}p_{A_2B_1})$, the most probable haplotypes and the weighted haplotypes for loci A and B result in the following for the phase-unknown double heterozygote genotype $A_1B_2A_2B_1$:

$$\begin{aligned} \text{most probable haplotypes - } A_1B_2A_2B_1 &\rightarrow \begin{cases} A_1B_1/A_2B_2 & \text{if } w \text{ is larger than } 0.5 \\ A_1B_2/A_2B_1 & \text{otherwise} \end{cases} \\ \text{weighted haplotypes - } A_1B_2A_2B_1 &\rightarrow \begin{cases} A_1B_1/A_2B_2 & \text{with probability } w \\ A_1B_2/A_2B_1 & \text{with probability } 1 - w. \end{cases} \end{aligned}$$

Thus, if we let N and $N_{A_kB_kA_k'B_k'}$ be the sample size and the number of individuals with genotype $A_kB_kA_k'B_k'$ in either cases or controls, the estimated parameters for LD $\widehat{\Delta}_{LD}^m$ and $\widehat{\Delta}_{LD}^w$ can be denoted

$$\begin{aligned} \widehat{\Delta}_{A_1B_1}^m &\equiv \frac{1}{N} \left[N_{A_1A_1B_1B_1} + \frac{1}{2}N_{A_1A_1B_1B_2} + \frac{1}{2}N_{A_1A_2B_1B_1} + \frac{1}{2}N_{A_1A_2B_1B_2} I\left(w > \frac{1}{2}\right) \right] - \widehat{p}_{A_1} \widehat{p}_{B_1} \\ \text{and } \widehat{\Delta}_{A_1B_1}^w &\equiv \frac{1}{N} \left[N_{A_1A_1B_1B_1} + \frac{1}{2}N_{A_1A_1B_1B_2} + \frac{1}{2}N_{A_1A_2B_1B_1} + \frac{w}{2}N_{A_1A_2B_1B_2} \right] - \widehat{p}_{A_1} \widehat{p}_{B_1}. \end{aligned}$$

By defining $\widehat{\Delta}_{A_1B_1}^W = \widehat{\Delta}_{A_1B_1}^m$ if $W = I(w > 1/2)$ for the indicator function I , and $\widehat{\Delta}_{A_1B_1}^W = \widehat{\Delta}_{A_1B_1}^w$ if $W = w$, we can subsume both estimates in a general way, using W , as follows:

$$\widehat{\Delta}_{A_1B_1}^W \equiv \frac{1}{N} \left[N_{A_1A_1B_1B_1} + \frac{1}{2}N_{A_1A_1B_1B_2} + \frac{1}{2}N_{A_1A_2B_1B_1} + \frac{W}{2}N_{A_1A_2B_1B_2} \right] - \widehat{p}_{A_1} \widehat{p}_{B_1}.$$

To derive the variance and covariances with the estimators for I_a and I_{HWD} , we parameterize $\widehat{\Delta}_{A_1B_1}^W$ in terms of the random variables defined above, as follows:

$$\begin{aligned}\widehat{\Delta}_{A_1 B_1}^W &= \frac{1}{2N} \sum_{i=1}^N [2X_{i1}X_{i2}Y_{i1}Y_{i2} + (1-X_{i1})X_{i2}Y_{i1}Y_{i2} + X_{i1}(1-X_{i2})Y_{i1}Y_{i2} \\ &\quad + X_{i1}X_{i2}(1-Y_{i1})Y_{i2} + X_{i1}X_{i2}Y_{i1}(1-Y_{i2}) + W(1-X_{i1})X_{i2}(1-Y_{i1})Y_{i2} \\ &\quad + [WX_{i1}(1-X_{i2})(1-Y_{i1})Y_{i2} + W(1-X_{i1})X_{i2}Y_{i1}(1-Y_{i2}) + WX_{i1}(1-X_{i2})Y_{i1}(1-Y_{i2})] \\ &\quad - \widehat{p}_{A_1}\widehat{p}_{B_1}] \\ &= \widehat{\Delta}_{A_1 B_1} + \frac{1}{2N} \sum_{i=1}^N [[W(X_{i1}(1-X_{i2})(1-Y_{i1})Y_{i2} + (1-X_{i1})X_{i2}Y_{i1}(1-Y_{i2}))]] \\ &\quad - [(1-W)\{(1-X_{i1})X_{i2}(1-Y_{i1})Y_{i2} + X_{i1}(1-X_{i2})Y_{i1}(1-Y_{i2})\}],\end{aligned}$$

because $\widehat{\Delta}_{A_1 B_1} = \sum_{i=1}^N [X_{i1}Y_{i1} + X_{i2}Y_{i2}] / 2N - \widehat{p}_{A_1}\widehat{p}_{B_1}$ and we obtain the results under Hardy-Weinberg equilibrium (HWE) using the delta method in a way analogous to that in Won and Elston [Won and Elston, 2008]:

$$\begin{aligned}E(\widehat{\Delta}_{A_1 B_1}^W) &= \Delta_{A_1 B_1} - (p_{A_1 B_1} + \Delta_{A_1 B_1}) \{ (1-p_{A_1})(1-p_{B_1}) + \Delta_{A_1 B_1} \} + W [2p_{A_1}p_{B_1} \times \\ &\quad [(1-p_{A_1})(1-p_{B_1}) + \Delta_{A_1 B_1} \{ (1-2p_{A_1})(1-2p_{B_1}) + 2\Delta_{A_1 B_1} \}] - \frac{\Delta_{A_1 B_1}}{2N}] \\ \text{var}(\widehat{\Delta}_{A_1 B_1}^W) &= \frac{1}{N} \left[(1-2W) \left\{ -\frac{1}{4}p_{A_1}q_{A_1}p_{B_1}q_{B_1} (4p_{A_1}q_{A_1}p_{B_1}q_{B_1} + 2b_{A_1}b_{B_1} - 1) + \Delta_{A_1 B_1} (-2p_{A_1}^2p_{B_1}^2 (2q_{A_1}q_{B_1} + q_{A_1} + q_{B_1})) \right\} \right. \\ &\quad + \left(2p_{A_1}p_{B_1} (p_{A_1}q_{A_1} + p_{B_1}q_{B_1} - 2q_{A_1}q_{B_1}) + \frac{1}{8} (-2(p_{A_1} - p_{B_1})^2 + 6(p_{A_1}q_{A_1} + p_{B_1}q_{B_1}) - 1) \right) + \Delta_{A_1 B_1}^2 (p_{A_1}q_{A_1} + p_{B_1}q_{B_1} \\ &\quad - 6p_{A_1}q_{A_1}p_{B_1}q_{B_1}) + \Delta_{A_1 B_1}^3 (1 - b_{A_1}b_{B_1}) - \Delta_{A_1 B_1}^4 + W \left(\frac{1}{2}p_{A_1}q_{A_1}p_{B_1}q_{B_1} (4p_{A_1}q_{A_1}p_{B_1}q_{B_1} - 1) + \frac{1}{4}\Delta_{A_1 B_1}b_{A_1}b_{B_1} (8p_{A_1}q_{A_1}p_{B_1}q_{B_1} \right. \\ &\quad \left. \left[((-1) + 2\Delta_{A_1 B_1}^2 (6p_{A_1}q_{A_1}p_{B_1}q_{B_1} - p_{A_1}q_{A_1} - p_{B_1}q_{B_1}) + 2\Delta_{A_1 B_1}^3 b_{A_1}b_{B_1} + 2\Delta_{A_1 B_1}^4) \right] + \frac{1}{8} (2p_{A_1}q_{A_1}p_{B_1}q_{B_1} + \Delta_{A_1 B_1}b_{A_1}b_{B_1}) \right) \left. \right] \\ \text{cov}(\widehat{p}_{A_1}, \widehat{\Delta}_{A_1 B_1}^W) &= \frac{1}{2N} (1-2p_{A_1}) [\Delta_{A_1 B_1} + 2Wp_{A_1 B_2}p_{A_2 B_1} - 2(1-W)p_{A_1 B_1}p_{A_2 B_2}] + O\left(\frac{1}{N^2}\right) \\ \text{cov}(\widehat{d}_A, \widehat{\Delta}_{A_1 B_1}^W) &= \frac{1}{2N} p_{A_1} [Wp_{A_1 B_2}p_{A_2 B_1} - (1-W)p_{A_1 B_1}p_{A_2 B_2}] + O\left(\frac{1}{N^2}\right).\end{aligned}$$

Next, the composite LD can be estimated as follows [Weir, 1996]:

$$\widehat{\Delta}_{A_1 B_1}^C \equiv \frac{1}{N} \left[2N_{A_1 A_1 B_1 B_1} + N_{A_1 A_1 B_1 B_2} + N_{A_1 A_2 B_1 B_1} + \frac{1}{2}N_{A_1 A_2 B_1 B_2} \right] - 2\widehat{p}_{A_1}\widehat{p}_{B_1}.$$

To derive its variance and relevant covariances, we parameterize it in terms of the random variables defined above as follows:

$$\begin{aligned}\widehat{\Delta}_{A_1 B_1}^C &= \frac{1}{2N} \sum_{i=1}^N [4X_{i1}X_{i2}Y_{i1}Y_{i2} + 2(1-X_{i1})X_{i2}Y_{i1}Y_{i2} + 2X_{i1}(1-X_{i2})Y_{i1}Y_{i2} \\ &\quad + 2X_{i1}X_{i2}(1-Y_{i1})Y_{i2} + 2X_{i1}X_{i2}Y_{i1}(1-Y_{i2}) + (1-X_{i1})X_{i2}(1-Y_{i1})Y_{i2} \\ &\quad + X_{i1}(1-X_{i2})(1-Y_{i1})Y_{i2} + (1-X_{i1})X_{i2}Y_{i1}(1-Y_{i2}) + X_{i1}(1-X_{i2})Y_{i1}(1-Y_{i2})] - 2\widehat{p}_{A_1}\widehat{p}_{B_1} \\ &= \widehat{\Delta}_{A_1 B_1} + \frac{1}{2N} \sum_{i=1}^N [X_{i1}Y_{i2} + X_{i2}Y_{i1}] - \widehat{p}_{A_1}\widehat{p}_{B_1}.\end{aligned}$$

With these results, we find that the expectations, variances [Weir, 1996] and covariances under HWE in either cases or controls are as follows:

$$E(\widehat{\Delta}_{A_1 B_1}^C) = \Delta_{A_1 B_1} - \frac{\Delta_{A_1 B_1}}{N},$$

$$\text{var}(\widehat{\Delta}_{A_1 B_1}^C) = \frac{1}{N} \left[p_{A_1} (1 - p_{A_1}) + p_{B_1} (1 - p_{B_1}) + \frac{1}{2} (1 - 2p_{A_1}) (1 - 2p_{B_1}) \Delta_{A_1 B_1} \right].$$

$$\text{cov}(\widehat{p}_{A_1}, \widehat{\Delta}_{A_1 B_1}^C) = \frac{\Delta_{A_1 B_1} (1 - 2p_{A_1})}{2N} + O\left(\frac{1}{N^2}\right), \text{ and}$$

$$\text{cov}(\widehat{d}_A, \widehat{\Delta}_{A_1 B_1}^C) = \frac{1}{N} \Delta_{A_1 B_1} p_{A_1} (1 - p_{A_1}) + O\left(\frac{1}{N^2}\right).$$

(see the Appendix for HWD). These show that the expectation and covariances of $\widehat{\Delta}_{A_1 B_1}^C$, but not its variance, are approximately equal to those of $\widehat{\Delta}_{A_1 B_1}$ under HWE. Also, the equivalence of $\Delta_{A_1 B_1}^C$ to $\Delta_{A_1 B_1}$ under HWE guarantees that

$$\text{var}(\widehat{\Delta}_{A_1 B_1}^C) = \frac{1}{N} \left[p_{A_1} (1 - p_{A_1}) p_{B_1} (1 - p_{B_1}) + \frac{1}{2} (1 - 2p_{A_1}) (1 - 2p_{B_1}) \Delta_{A_1 B_1}^C \right] \text{ and}$$

$$\text{cov}(\widehat{p}_{A_1}, \widehat{\Delta}_{A_1 B_1}^C) = \frac{\Delta_{A_1 B_1}^C (1 - 2p_{A_1})}{2N} + O\left(\frac{1}{N^2}\right).$$

With these results, we can consider the following three LD contrast test statistics for association analysis in case-control studies:

$$S_{LD}^m = \frac{\widehat{\Delta}_{A_1 B_1 | \text{case}}^m - \widehat{\Delta}_{A_1 B_1 | \text{control}}^m}{\sqrt{\text{var}(\widehat{\Delta}_{A_1 B_1 | \text{case}}^m - \widehat{\Delta}_{A_1 B_1 | \text{control}}^m)}} \quad S_{LD}^w = \frac{\widehat{\Delta}_{A_1 B_1 | \text{case}}^w - \widehat{\Delta}_{A_1 B_1 | \text{control}}^w}{\sqrt{\text{var}(\widehat{\Delta}_{A_1 B_1 | \text{case}}^w - \widehat{\Delta}_{A_1 B_1 | \text{control}}^w)}} \\ S_{LD}^c = \frac{\widehat{\Delta}_{A_1 B_1 | \text{case}}^c - \widehat{\Delta}_{A_1 B_1 | \text{control}}^c}{\sqrt{\text{var}(\widehat{\Delta}_{A_1 B_1 | \text{case}}^c - \widehat{\Delta}_{A_1 B_1 | \text{control}}^c)}}.$$

Quantification of the parameters for LD

To elucidate the relative efficiency of the parameters for LD, the information in I_{LD}^m , I_{LD}^w and I_{LD}^c needs to be quantified and compared. If we assume HWE in the population and respectively let ϕ and ϕ_l be the disease prevalence and the penetrance of genotype l , the parameters for I_a , I_{HWD} and I_{LD} in cases and controls are as follows [Nielsen et al., 1998; Won and Elston, 2008]:

$$p_{A_1 | \text{case}} = p_{A_1} + \frac{\Delta_{A_1 D_1}}{\phi} \left[p_{D_1} (\phi_{D_1 D_1} - \phi_{D_1 D_2}) + p_{D_2} (\phi_{D_1 D_2} - \phi_{D_2 D_2}) \right]$$

$$p_{A_1 | \text{control}} = p_{A_1} + \frac{\Delta_{A_1 D_1}}{1 - \phi} \left[p_{D_1} (\phi_{D_1 D_1} - \phi_{D_1 D_2}) + p_{D_2} (\phi_{D_1 D_2} - \phi_{D_2 D_2}) \right]$$

$$d_{A | \text{case}} = p_{A_1 A_1 | \text{case}} - (p_{A_1 | \text{case}})^2 = \frac{\Delta_{A_1 D_1}^2}{\phi} (\phi_{D_1 D_1} \phi_{D_2 D_2} - \phi_{D_1 D_2}^2)$$

$$d_{A | \text{control}} = p_{A_1 A_1 | \text{control}} - (p_{A_1 | \text{control}})^2 = \frac{\Delta_{A_1 D_1}^2}{(1 - \phi)^2} \left((1 - \phi_{D_1 D_1}) (1 - \phi_{D_2 D_2}) - (1 - \phi_{D_1 D_2})^2 \right)$$

$$\begin{aligned}\Delta_{A_1 B_1 | \text{case}} &= \Delta_{A_1 B_1} + \frac{1}{\phi} [p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})] \\ &\times \left[\Delta_{A_1 D B_1} - \frac{1}{\phi} (p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})) \Delta_{A_1 D} \Delta_{D B_1} \right] \\ \Delta_{A_1 B_1 | \text{control}} &= \Delta_{A_1 B_1} + \frac{1}{1-\phi} [p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})] \\ &\times \left[\Delta_{A_1 D B_1} - \frac{1}{1-\phi} (p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})) \Delta_{A_1 D} \Delta_{D B_1} \right].\end{aligned}$$

The quantification of $\Delta_{A_1 B_1}^m$ and $\Delta_{A_1 B_1}^w$ can be generalized using $\Delta_{A_1 B_1}^W$:

$$\begin{aligned}\Delta_{A_1 B_1 | \text{case}}^W &= (2W - 1) p_{A_1 | \text{case}} p_{A_2 | \text{case}} p_{B_1 | \text{case}} p_{B_2 | \text{case}} + \Delta_{A_1 B_1 | \text{case}} \left\{ 1 + (W - 1) (p_{A_1 | \text{case}} p_{B_1 | \text{case}}) \right\} \\ &\left\{ + p_{A_2 | \text{case}} p_{B_2 | \text{case}} - W (p_{A_1 | \text{case}} p_{B_2 | \text{case}} + p_{A_2 | \text{case}} p_{B_1 | \text{case}}) \right\} + (2W - 1) \Delta_{A_1 B_1 | \text{case}}^2,\end{aligned}$$

and $\Delta_{A_1 B_1 | \text{control}}^W$ can also be quantified analogously.

The quantification of $\Delta_{A_1 B_1}^C$ requires the joint frequencies of alleles A_1 and B_1 in two different gametes because $\Delta_{A_1 B_1}^C$ is defined as $\Delta_{A_1 B_1}^C = p_{A_1 B_1} + p_{A_1 | B_1} - 2 p_{A_1} p_{B_1}$. Under HWE in the population we have:

$$\begin{aligned}p_{A_1 | B_1 | \text{case}} &= \frac{1}{\phi} \sum_{j \in \{DD, Dd, dd\}} \left[P \left(\begin{matrix} A_1 \\ B_1 \end{matrix} \middle| \begin{matrix} A_1 \\ B_1 \end{matrix}, j, \text{affected} \right) + \frac{1}{2} \left\{ P \left(\begin{matrix} A_1 \\ B_2 \end{matrix} \middle| \begin{matrix} A_1 \\ B_1 \end{matrix}, j, \text{affected} \right) \right\} \right] \\ &\left[\left\{ + P \left(\begin{matrix} A_1 \\ B_1 \end{matrix} \middle| \begin{matrix} A_2 \\ B_1 \end{matrix}, j, \text{affected} \right) + P \left(\begin{matrix} A_1 \\ B_2 \end{matrix} \middle| \begin{matrix} A_2 \\ B_1 \end{matrix}, j, \text{affected} \right) \right\} \right] \\ &= \frac{1}{\phi} \left[P(A_1 D_1) P(D_1 B_1) \phi_{D_1 D_1} + \{ P(A_1 D_1) P(D_2 B_1) + P(A_1 D_2) P(D_1 B_1) \} \phi_{D_1 D_2} \right] \\ &\quad + P(A_1 D_2) \left[P(D_2 B_1) \phi_{D_2 D_2} \right] \\ &= p_{A_1} p_{B_1} + \frac{1}{\phi} \left[(p_{A_1} \Delta_{D_1 B_1} + p_{B_1} \Delta_{A_1 D_1}) \{ p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd}) \} \right. \\ &\quad \left. + \Delta_{A_1 D_1} \Delta_{D_1 B_1} (\phi_{D_1 D_1} - 2\phi_{D_1 D_2} + \phi_{D_2 D_2}) \right],\end{aligned}$$

where the derivation is based on the fact that $p_{A_1 | B_1} = p_{A_1 B_1} + \frac{1}{2} (p_{A_1 B_2} + p_{A_2 B_1} + p_{A_2 B_2})$. Thus, under HWE the composite LD between markers in cases is

$$\begin{aligned}\Delta_{A_1 B_1 | \text{case}}^C &= \Delta_{A_1 B_1 | \text{case}} + p_{A_1 | B_1 | \text{case}} - p_{A_1 | \text{case}} p_{B_1 | \text{case}} \\ &= \Delta_{A_1 B_1} + \frac{\Delta_{A_1 D_1} \Delta_{D_1 B_1}}{\phi} [\phi_{DD} - 2\phi_{Dd} + \phi_{dd}] + \frac{1}{\phi} [p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})] \\ &\quad \times \left[\Delta_{A_1 D B_1} - \frac{2}{\phi} (p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})) \Delta_{A_1 D} \Delta_{D B_1} \right].\end{aligned}$$

Analogously, for controls, we have:

$$\begin{aligned}\Delta_{A_1 B_1 | \text{control}}^C &= \Delta_{A_1 B_1 | \text{control}} + p_{A_1 | B_1 | \text{control}} - p_{A_1 | \text{control}} p_{B_1 | \text{control}} \\ &= \Delta_{A_1 B_1} + \frac{\Delta_{A_1 D_1} \Delta_{D_1 B_1}}{1-\phi} [\phi_{DD} - 2\phi_{Dd} + \phi_{dd}] + \frac{1}{1-\phi} [p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})] \\ &\quad \times \left[\Delta_{A_1 D B_1} - \frac{2}{1-\phi} (p_D (\phi_{DD} - \phi_{Dd}) + p_d (\phi_{Dd} - \phi_{dd})) \Delta_{A_1 D} \Delta_{D B_1} \right].\end{aligned}$$

Results

Simulations using real marker data

The cystic fibrosis transmembrane conductance regulator (CFTR) is located across 200 Kb in region q31.2 on the long arm of chromosome 7. From the HapMap database, we downloaded the marker data of the CFTR haplotype in the CEU (CEPH-Utah resident) sample. We randomly selected, to generate separate genotypes for each of 4,000 persons (2,000 representing cases and 2,000 representing controls), 11 SNPs whose minor allele frequencies are larger than 0.1 and used the SNP located in the middle (with allele frequency ranging between 0.1 and 0.5) as a causal SNP for the 2,000 cases, the others being used as markers. To study empirical type I error, we assume homozygous and heterozygote disease genotype relative risks λ_2 and λ_1 equal to 1, and randomly generated a binary trait. To study empirical power, we assumed $\lambda_2=1.4$, λ_1 being determined by the disease mode of

inheritance: $\lambda_1 = 1$ for a recessive disease, $\lambda_1 = \lambda_2$ for a dominant disease, $\lambda_1 = \frac{\lambda_2+1}{2}$ for an additive disease and $\lambda_1 = \sqrt{\lambda_2}$ for a multiplicative disease. We considered two marker densities: (i) sparse markers, the consecutive markers being separated by about 30 Kb, and (ii) dense markers, separated by about 5Kb. For each situation, the empirical power and size was calculated from 1,000 replicates of 2,000 cases and 2,000 controls. Thus, in all, $4,000 \times 1,000 = 4,000,000$ separate genotypes were generated.

In these simulations, the statistics, S_a and S_{HW} were calculated at the marker closest to the causal SNP for the empirical power. S_{LD} requires information about phase, and so the haplotypes for each individual were first estimated from the 10 markers with the default options of the program DECIPHER in S.A.G.E.¹ and then the most probable pair of haplotypes assumed to be the true haplotypes. Table 1 shows that for sparse markers, whereas the type I errors for S_a and S_{HWD} are little inflated or deflated, S_{LD} can be too conservative. Also, it shows that S_a is the most powerful statistic and the others are not very informative. Contrary to the results for sparse markers, Table 2 shows that, for dense markers, not only do all three statistics preserve type I error relatively well (though S_{LD}^C is a little liberal), but also S_{LD} yields the most informative test. Thus, when markers are not dense enough, I_{LD} can be almost uninformative, and assigning the most probable haplotypes as true haplotypes results in inflated (or sometimes deflated) type I error. If the markers are dense, on the other hand, we can assume the most probable haplotypes are the true haplotypes and I_{LD} can be more informative than I_a . It has in fact been shown that the power of S_{LD} depends on high multi-locus LD between marker and disease alleles [Won and Elston, 2008], which explains why I_{LD} can be more informative than I_a when the markers are dense.

Simulations using simulated marker data

From the previous simulations, we confirmed that for sparse markers the type I error of S_{LD} can be inflated if the most probable haplotypes are assumed to be true haplotypes. We now examine with simulations both the validity and efficiency of the three LD contrast tests modified for phase uncertainty: S_{LD}^m , S_{LD}^w and S_{LD}^c . In particular the two-marker haplotype frequencies for S_{LD}^m and S_{LD}^w are estimated by the EM algorithm [Excoffier and Slatkin, 1995]. The disease prevalence and the disease allele frequency are assumed to be 0.15 and 0.1, respectively. We used $\lambda_2 = 1.4$, λ_1 being determined by the disease mode of inheritance. For the empirical type I error, we assume LE between marker and disease alleles and, for the empirical power, we considered three different LD structures: (1) $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.023$, $\Delta_{A_1B_1} = 0.06$ and $\Delta_{A_1D_1B_1} = 0.013$; (2) $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.047$, $\Delta_{A_1B_1} = 0.06$

and $\Delta_{A_1D_1B_1} = 0.03$; and (3) $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.068$, $\Delta_{A_1B_1} = 0.06$ and $\Delta_{A_1D_1B_1} = 0.049$. These values represent three increasing levels of three-locus LD.

Figures 2 and 3 show the empirical type I error from 20,000 replicates of 10,000 cases and 10,000 controls as a function of LD between markers when p_{A_1} and p_{B_1} are 0.2 and 0.3, respectively. For w , we considered two different situations: (i) the haplotype frequencies in cases and controls are the same, so that the weights, w , are equal for cases and controls, and (ii) the haplotype frequencies in cases and controls are different. The results show that estimation of the weight, w , can result in preserved type I errors for S_{LD}^m and S_{LD}^w under case (i) but results in inflated type I error under case (ii), while S_{LD}^c always preserves the type I error well. However, when the absolute values of Lewontin's D' [Lewontin, 1964] between markers are high, in both situations the inflation of type I error is negligible. Thus we conclude that type I error is preserved well as long as the same weights are used for cases and controls.

Figures 4, 5 and 6 show the average empirical power from 20,000 replicates for three different LD structures as a function of sample size; for w in S_{LD}^m and S_{LD}^w we assume that the haplotype frequencies are the same in cases and controls. The sample sizes for cases and controls are equal. From the results, we can confirm that the LD contrast test (S_{LD}^m , S_{LD}^w and S_{LD}^c) can be more powerful than S_a with LD structure (3), because there is high three-locus LD between marker and disease alleles. Also, as was shown before, S_{LD}^c is better than S_{LD}^m and S_{LD}^w for a recessive disease. Otherwise, S_{LD}^c loses power because its variance is approximately twice as large as the variance of S_{LD}^m or S_{LD}^w .

Figure 7 shows, for p_{A_1} and $p_{B_1} = 0.2$, the analytical correlations of $\hat{p}_{A_1}$ and $\hat{d}_A$ with $\hat{\Delta}_{A_1B_1}^m$, $\hat{\Delta}_{A_1B_1}^w$ and $\hat{\Delta}_{A_1B_1}^c$ when the haplotype frequencies are known (panels **a** and **b**), and the empirical correlations of S_a and S_{HWD} with S_{LD}^m , S_{LD}^w and S_{LD}^c (panels **c** and **d**) obtained from 20,000 replicates of 10,000 cases and 10,000 controls under LE between marker and disease alleles. Overall, panels **a** and **b** are similar to panels **c** and **d**, respectively, even though haplotype frequencies are assumed to be known for the former and not for the latter. Figure 7 panels **c** and **d** also show that, for $D' > 0$, S_a is least correlated with S_{LD}^c and most with S_{LD}^m , while S_{HWD} is least correlated with S_{LD}^m and most with S_{LD}^c . In particular, the correlations of S_{HWD} with the parameters for LD are relatively small because HWE is assumed in the population. As a result, either S_{LD}^c or S_{LD}^w may then be appropriate for self-replication as long as they have comparable informativity.

Quantification

The above simulations confirmed that the methods for handling phase uncertainty preserve the type I error in certain situations. We now examine quantification of the standardized expected differences between cases and controls in order to analytically compare their efficiencies. If we let SED be the standardized expected difference for a single case and a single control, defined as

$$SED = \frac{\text{expectation of the difference statistic for one case and one control}}{\text{s.d. of the difference statistic for one case and one control}},$$

and if we let $\Phi(\cdot)$ and $Z_{1-\alpha/2}$ be respectively the cumulative standard normal distribution and the $1-\alpha/2$ quantile of the standard normal distribution, the power for N cases and N controls

is equal to $\Phi(SED\sqrt{N} - Z_{1-\alpha/2}) + \Phi(-SED\sqrt{N} - Z_{1-\alpha/2})$. The required sample size for given power and significance level can be plotted as a function of the SED [Won and Elston, 2008]. In Figures 8 and 9, we plot SED s as a function of LD for various modes of inheritance. Again, the disease allele frequency was assumed to be 0.1 and we used $\lambda_2 = 1.4$, λ_1 being determined by the disease mode of inheritance. We assume the disease prevalence is 0.15 and the minor allele frequencies for the markers A and B are 0.2. Figure 8 shows the SED as a function of $\Delta_{A_1D_1}$, when $\Delta_{A_1D_1B_1} = 0.25\Delta_{A_1D_1}$ and $\Delta_{B_1D_1} = 0.2\Delta_{A_1D_1}$. These results show that S_{LD}^C is less powerful than S_{LD} except in the case of a recessive disease. The expectation of the statistic for I_{LD}^C is slightly larger than that of I_{LD} , but the variance is much larger; if $\Delta_{A_1B_1}$ is small, the variance of the composite LD estimator is approximately twice as large as the variance of the LD estimator based on phase-known data. When the genetic effects are large, the variance of I_{LD}^C is much larger than that of I_{LD} , while the expected differences for I_{LD}^C and I_{LD} are similar, which explains the higher SED of I_{LD} . In addition, it can be seen in Figure 8 how the disease mode of inheritance affects the SED s of the three statistics. For a recessive model, the estimators for I_a and I_{HWD} are relatively better, whereas the estimators for I_a and I_{LD} (I_{LD}^m , I_{LD}^w and I_{LD}^C) are better for the other disease models (see Won and Elston¹⁵ for further details). Figure 9 shows the SED as a function of $\Delta_{A_1B_1}$, when $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.068$ and $\Delta_{A_1D_1B_1} = 0.049$. As expected, the results show that the SED of I_{LD}^C is largest for a recessive disease and the SED of I_{LD} is largest for the other disease models. The SED s of I_{LD} and I_{LD}^C are inversely proportional to the magnitude of $\Delta_{A_1B_1}$ because the variances increase with $\Delta_{A_1B_1}$, though the expected difference does not depend on $\Delta_{A_1B_1}$. In addition, the SED s for I_{LD}^m and I_{LD}^w are similar to that for I_{LD} when $\Delta_{A_1B_1}$ is high, while the SED s of I_{LD}^m and I_{LD}^w are almost negligible when $\Delta_{A_1B_1}$ is negative. As a result, we can conclude that I_{LD}^C is generally valid and efficient but, when $\Delta_{A_1B_1}$ is positive and the disease mode of inheritance is not recessive, either I_{LD}^m or I_{LD}^w is better when we have dense markers.

HWD in the population and type I error

Table 3 shows the empirical type I error according to whether there is HWE or HWD in the population. p_{A_1} , p_{B_1} and p_{D_1} are assumed to be 0.2, 0.2 and 0.1, respectively, λ_2 is 1.4 and λ_1 is calculated according to the disease mode of inheritance. The disease prevalence is 0.15 and LE is assumed between marker and disease alleles. In all cases we assumed the parameter for haplotype-level HWD for the three loci (A , B , and D) is 6.4×10^{-5} , which is the maximal amount possible in this situation. Two types of variances for S_a , S_{HWD} and S_{LD}^C are considered: (i) calculated allowing for HWD, and (ii) calculated assuming HWE. Table 3, based on 20,000 replicates each with 10,000 cases and 10,000 controls, shows that while S_a , S_{HWD} and S_{LD}^C preserve the type I errors at the significance levels 0.05 and 0.01 for all disease modes of inheritance in situation (i), the empirical type I errors are inflated in situation (ii). Also, it appears that the disease mode of inheritance and the type of statistic affect the level of inflation. S_a is inflated only for a dominant disease, while S_{HWD} and S_{LD}^C are inflated for all disease modes of inheritance. In particular, S_{LD}^C is the most sensitive to HWD, which indicates that HWE in the population should be confirmed *a priori* for the LD contrast test.

Haplotype-level HWD under LE between marker and disease alleles can result in differences between cases and controls for three types of parameters. For example, the allele frequencies for cases and controls are

$$\begin{aligned}
 p_{A_1|case} &= \frac{1}{\phi} \left[(p_{A_1D_1}^2 + p_{A_1D_1} p_{A_2D_1} + d_{12}^{AD} + d_{14}^{AD}) \phi_{D_1D_1} + (2p_{A_1D_1} p_{A_1D_2} + p_{A_1D_1} p_{A_2D_2} + p_{A_1D_2} p_{A_2D_1}) \right. \\
 &\quad \left. [-2d_{12}^{AD} - d_{14}^{AD} - d_{23}^{AD}] \phi_{D_1D_2} + (p_{A_1D_2}^2 + p_{A_1D_2} p_{A_2D_2} + d_{12}^{AD} + d_{23}^{AD}) \phi_{D_2D_2} \right] \\
 &= p_{A_1} + \frac{1}{\phi} \left[(\Delta_{A_1D_1} p_{D_1} + d_{12}^{AD} + d_{14}^{AD}) (\phi_{D_1D_1} - \phi_{D_1D_2}) + (\Delta_{A_1D_1} p_{D_2} - d_{12}^{AD} - d_{23}^{AD}) (\phi_{D_1D_2} - \phi_{D_2D_2}) \right], \\
 p_{A_1|control} &= p_{A_1} \frac{1}{1-\phi} \left[(\Delta_{A_1D_1} p_{D_1} + d_{12}^{AD} + d_{14}^{AD}) (\phi_{D_1D_1} - \phi_{D_1D_2}) + (\Delta_{A_1D_1} p_{D_2} - d_{12}^{AD} - d_{23}^{AD}) (\phi_{D_1D_2} - \phi_{D_2D_2}) \right].
 \end{aligned}$$

These results indicate that the allele frequencies can be different between cases and controls under the null hypothesis $\Delta_{A_1D_1} = 0$, if there is haplotype-level HWD in the population. However, Table 3 shows that type I error is preserved well and its effect is negligible for most situations as long as we use the variances that allow for HWD.

Discussion

In addition to confirming previous results [Won and Elston, 2008], we have shown that, if markers are dense enough, the LD contrast test can be more informative than S_a . Also, we extended S_{LD} to S_{LD}^m , S_{LD}^w and S_{LD}^c , the last of which does not require any information about phase. With this extension, we made all the types of information feasible for testing association in a case-control study. S_{LD}^m and S_{LD}^w need estimates of the haplotype frequencies and our simulation shows that haplotype frequency estimation does not substantially inflate the type I error as long as the same weights are used for cases and controls. Otherwise, it greatly inflates type I error when the absolute LD between marker alleles is small, while type I error is preserved for high LD between markers. Our results also show that the variance of I_{LD}^c is approximately twice as large as the variance of I_{LD} .

The decomposition of the parameters available for a case-control study enable self-replication and the increase in power resulting from self-replication can be understood in terms of methods for combining p-values. The most powerful method has been derived for when the SEDs are known and this method can be extended to statistics that are correlated¹⁶. Also, it should be remembered that among the possible choices for phase uncertainty S_{LD}^m is the least appropriate when the markers are not dense because it is then the most correlated with the statistics that use the other types of information (see Figure 7). If the disease mode of inheritance is known, in the case of a recessive disease we suggest combining the p-values from S_a , S_{HWD} and S_{LD}^c . For a dominant disease, we should combine the p-values from S_a , S_{HWD} and S_{LD}^w (or S_{LD}^m) when markers are dense, and the p-values from S_a , S_{HWD} and S_{LD}^c when the markers are not dense. For multiplicative and additive disease modes of inheritance, we should consider S_a and S_{LD}^m (or S_{LD}^w) when the markers are dense, and S_a and S_{LD}^c when the markers are not dense. If the disease mode of inheritance is unknown, we should first decide on the mode of inheritance using S_{HWD} [Wittke-Thompson et al., 2005; Zheng and Ng, 2008], and then combine S_a and S_{LD}^m (or S_{LD}^w), or S_a and S_{LD}^c according to the disease mode of inheritance and marker density (Figure 10 shows the SED as a function of λ_1).

However, in spite of the efficiency and validity of the proposed method, there are still problems that need attention. First, the suggested strategy that determines the disease mode of inheritance with S_{HWD} does not use S_{HWD} to indicate association *per se*. As has been shown, the most powerful method is attainable only when the SEDs of the statistics are known; the SED for I_{HWD} needs information about the disease mode of inheritance. The new method of determining the disease mode of inheritance with type I error preserved requires

further investigation. Second, we need to have estimates of the covariances to combine the three types of information if HWE in the population is violated. If HWE in the population may be assumed, the covariances can be estimated from the equivalence of the composite LD and LD. However, under HWD the equivalence is violated and even more parameters need to be estimated that require information about phase.

In summary, even though there have now been many investigations using GWA, except for genetic mapping based on linkage disequilibrium units [Collins and Lau, 2008], they have so far considered only the statistic based on I_a for an initial scan, which can be powerless in some cases. Also, application to real marker data indicates that the LD contrast test can be very informative in some situations, so that combining p-values from S_a and S_{LD}^c can increase the statistical power of GWA studies. The proposed method is optimal and efficient from the viewpoint of using all the information available.

Acknowledgments

This work was supported in part by a U.S. Public Health Service Resource grant (RR03655) from the National Center for Research Resources, Research grant (GM28356) from the National Institute of General Medical Sciences and Cancer Center Support Grant (P30CAD43703) from the National Cancer Institute.

Appendix

1. Notation

Before we derive the expectations, variances and covariances, we define the following notation that is used in Weir [Weir, 1996]:

$$\begin{aligned}\Delta_{A_1/B_1} &\equiv p_{A_1/B_1} - p_{A_1} p_{B_1} = p_{A_1} p_{B_1} + d_{14}^{AB} - d_{23}^{AB} - p_{A_1} p_{B_1} = d_{14}^{AB} - d_{23}^{AB} \\ d_{A_1 A_1 B_1} &\equiv p_{A_1 B_1}^{A_1} + \frac{1}{2} p_{A_1 B_1}^{A_1} - p_{A_1} \Delta_{A_1 B_1}^C - p_{B_1} d_A - p_{A_1}^2 p_{B_1} \\ d_{A_1 B_1 B_1} &\equiv p_{A_1 B_1}^{A_1} + \frac{1}{2} p_{A_1 B_1}^{A_1} - p_{B_1} \Delta_{A_1 B_1}^C - p_{A_1} d_B - p_{A_1} p_{B_1}^2 \\ &= d_{12}^{AB} (1 - p_{A_1}) + d_{14}^{AB} (1 - p_{A_1} - p_{B_1}) + d_{23}^{AB} (p_{B_1} - p_{A_1}) - d_{34}^{AB} p_{A_1} \\ d_{A_1 B_1}^{A_1 B_1} &\equiv p_{A_1 B_1}^{A_1} - 2p_{A_1} d_{A_1 B_1 B_1} - 2p_{B_1} d_{A_1 A_1 B_1} - 2p_{A_1} p_{B_1} \Delta_{A_1 B_1}^C - p_{A_1}^2 d_B - p_{B_1}^2 d_A \\ &\quad - \Delta_{A_1 B_1}^2 - \Delta_{A_1/B_1}^2 - d_A d_B - p_{A_1}^2 p_{B_1}^2 \\ &= d_{12}^{AB} + d_{13}^{AB} + d_{14}^{AB} + 2p_{A_1} d_{A_1 B_1 B_1} - 2p_{B_1} d_{A_1 A_1 B_1} - 2p_{A_1} p_{B_1} \Delta_{A_1 B_1}^C \\ &\quad - p_{A_1}^2 d_B - p_{B_1}^2 d_A - \Delta_{A_1/B_1}^2 - d_A d_B,\end{aligned}$$

where it should be noted that $\Delta_{A_1 B_1}^C = \Delta_{A_1 B_1} + \Delta_{A_1/B_1}$. Under HWE, $\Delta_{A_1/B_1} = d_{A_1 A_1 B_1} = d_{A_1 B_1 B_1} = 0$

because $d_{A_1 A_1 B_1} = d_{13}^{AB} (1 - p_{B_1}) + d_{14}^{AB} (1 - p_{A_1} - p_{B_1}) + d_{23}^{AB} (p_{A_1} - p_{B_1}) - d_{24}^{AB} p_{B_1}$,

$d_{A_1 B_1 B_1} = d_{12}^{AB} (1 - p_{A_1}) + d_{14}^{AB} (1 - p_{A_1} - p_{B_1}) + d_{23}^{AB} (p_{B_1} - p_{A_1}) - d_{34}^{AB} p_{A_1}$.

2. Results under HWD

The variances of $\widehat{\Delta}_{A_1 B_1}$ and $\widehat{\Delta}_{A_1 B_1}^C$ were derived as follows [Weir, 1996]:

$$\begin{aligned} \text{var} \left(\widehat{\Delta}_{A_1 B_1} \right) &= \frac{1}{2N} \left[p_{A_1} (1 - p_{A_1}) p_{B_1} (1 - p_{B_1}) + (1 - 2p_{A_1})(1 - 2p_{B_1}) \Delta_{A_1 B_1} - \Delta_{A_1 B_1}^2 \right] \\ &\quad + d_A d_B + \Delta_{A_1/B_1}^2 + d_{A_1 B_1}^2 \\ \text{var} \left(\widehat{\Delta}_{A_1 B_1}^C \right) &= \frac{1}{N} \left[p_{A_1 B_1}^2 + p_{A_1} p_{B_1} (1 - p_{A_1} - p_{B_1}) + p_{A_1} (1 - 2p_{A_1}) d_B + p_{B_1} (1 - 2p_{B_1}) d_A + (1 - 4p_{A_1}) d_{A_1 B_1} \right] \\ &\quad + (1 - 4p_{B_1}) d_{A_1 A_1 B_1} + \frac{1}{2} (1 - 2p_{A_1} - 2p_{B_1}) \Delta_{A_1 B_1}^C - \Delta_{A_1 B_1}^C{}^2 \end{aligned}$$

With results from Won and Elston [Won and Elston, 2008], we have

$$\begin{aligned} E \left(\widehat{\Delta}_{A_1 B_1} \right) &= \Delta_{A_1 B_1} + p_{A_1 B_1} - \frac{1}{2N^2} E \left(\sum_{i=1}^N X_{i1} \right) \left(\sum_{i=1}^N Y_{i1} \right) - \frac{1}{2N^2} E \left(\sum_{i=1}^N X_{i1} \right) \left(\sum_{i=1}^N Y_{i2} \right) \\ &= \Delta_{A_1 B_1} - \frac{\Delta_{A_1 B_1} + d_{14}^{AB} - d_{23}^{AB}}{2N}, \\ E \left(\widehat{\Delta}_{A_1 B_1}^C \right) &= E \left(\widehat{\Delta}_{A_1 B_1} + \frac{1}{2N} \sum_{i=1}^N [X_{i1} Y_{i2} + X_{i2} Y_{i1}] - \widehat{p}_{A_1} \widehat{p}_{B_1} \right) \\ &= \Delta_{A_1 B_1} + d_{14}^{AB} - d_{23}^{AB} - \frac{\Delta_{A_1 B_1} + d_{14}^{AB} - d_{23}^{AB}}{N}, \\ \text{cov} \left(\widehat{\Delta}_{A_1}^C, \widehat{\Delta}_{A_1 B_1}^C \right) &= \frac{1}{2N} \left[-2p_{B_1} d_A + (d_{14}^{AB} - d_{23}^{AB}) + 2d_{13}^{AB} + 2d_{14}^{AB} - 4p_{A_1} (d_{14}^{AB} - d_{23}^{AB}) \right] \\ &\quad + \Delta_{A_1 B_1} (1 - 2p_{A_1}) + O \left(\frac{1}{N^2} \right), \\ \text{cov} \left(\widehat{d}_A, \widehat{\Delta}_{A_1 B_1}^C \right) &= \frac{1}{N} \left[-2p_{B_1} d_A (1 - 2p_{A_1}) + 2(d_{13}^{AB} + d_{14}^{AB})(1 - 2p_{A_1}) + (d_{14}^{AB} - d_{23}^{AB})(-p_{A_1} + 3p_{A_1}^2 - d_A) \right] \\ &\quad + \Delta_{A_1 B_1} (p_{A_1} - p_{A_1}^2 - d_A) + O \left(\frac{1}{N^2} \right). \end{aligned}$$

REFERENCES

- Collins A, Lau W. CHROMOSCAN: genome-wide association using a linkage disequilibrium map. *J Hum Genet.* 2008; 53:121–126. [PubMed: 18043862]
- Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 1995; 12:921–927. [PubMed: 7476138]
- Fisher, RA. *Statistical Methods for Research Workers.* 11th ed.. Oliver & Boyd; Edinburgh: 1950.
- Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics.* 1964; 49:49–67. [PubMed: 17248194]
- Liptak T. On the combination of independent tests. *Magyar Tud.Akad.Mat.Kutato' Int.Ko"zl.* 1958; 3:171–197.
- Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet.* 1998; 63:1531–1540. [PubMed: 9867708]
- Nielsen DM, Ehm MG, Zaykin DV, Weir BS. Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics.* 2004; 168:1029–1040. [PubMed: 15514073]
- S.A.G.E. 5.2. *Statistical Analysis for Genetic Epidemiology.* 2007. <http://darwin.cwru.edu/sage/>
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics.* 1997; 53:1253–1261. [PubMed: 9423247]
- Song K, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat in Med.* 2006; 25:105–126. [PubMed: 16220513]
- Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, DeMeo DL, Murphy A, Su J, Datta S, Rosenow C, et al. Genomic screening and replication using the same data set in family-based association testing. *Nat Genet.* 2005; 37:683–691. [PubMed: 15937480]
- Wang T, Zhu W, Elston RC. Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am J Hum Genet.* 2007; 80:911–920. [PubMed: 17436245]

- Weir, BS. Genetic Data Analysis II. Sinauer Associates; Sunderland, MA: 1996.
- Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet.* 2005; 76:967–986. [PubMed: 15834813]
- Won S, Elston RC. The power of independent types of genetic information to detect association in a case-control study design. *Genet Epidemiol.* 2008 In press.
- Won S, Morris N, Lu Q, Elston RC. Choosing optimal method for combining p-values. *Stat in Med.* 2008 Submitted.
- Zaykin DV. Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet Epidemiol.* 2004; 27:252–257. [PubMed: 15389931]
- Zaykin DV, Meng Z, Ehm MG. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am J Hum Genet.* 2006; 78:737–746. [PubMed: 16642430]
- Zheng G, Song K, Elston RC. Adaptive two-stage analysis of genetic association in case-control designs. *Hum Hered.* 2007; 63:175–186. [PubMed: 17310127]
- Zheng G, Ng HK. Genetic model selection in two-phase analysis for case control association studies. *Biostatistics.* 2008; 9:391–399. [PubMed: 18003629]

Figure 1. Rejection regions

the rejection regions are shown for the significance level 0.00001 in the negative log-scales when we have two independent p-values, u_1 and u_2 ; (a) when we use either of them alone (here illustrated for u_1), (b) self-replication, [Van Steen et al., 2005; Zheng et al., 2007] (c) Fisher's method of combining p-values, (d) Liptak's method of combining p-values [Liptak, 1958].

Figure 2. Empirical type I error with simulated marker data

p_{A1} and p_{B1} are 0.2; D' , Lewontin's D'^5 ; average of 20,000 replicates of 10,000 cases and 10,000 controls; significance level 0.05 with LE between marker and disease alleles; $p_{D1} = 0.15$, $\phi = 0.1$ and $\lambda_2 = 1.4$; S_{LD}^m and S_{LD}^w are calculated under the assumption that the haplotype frequencies in cases and controls are the same, and S_{LD}^m and S_{LD}^w under the assumption that the haplotype frequencies in cases and controls are different. The lines for S_{LD}^m , S_{LD}^w and S_{LD}^c are virtually identical around the nominal significance level.

Figure 3. Empirical type I error with simulated marker data

p_{A1} and p_{B1} are 0.3; D' , Lewontin's D'^5 ; average of 20,000 replicates of 10,000 cases and 10,000 controls; significance level 0.05 with LE between marker and disease alleles; $p_{D1} = 0.15$, $\phi = 0.1$ and $\lambda_2 = 1.4$; S_{LD}^m and S_{LD}^w are calculated under the assumption that the haplotype frequencies in cases and controls are the same, and $S_{LD}^{'m}$ and $S_{LD}^{'w}$ under the assumption that the haplotype frequencies in cases and controls are different. The lines for S_{LD}^m , S_{LD}^w and S_{LD}^c are virtually identical around the nominal significance level.

Figure 4. Empirical power when $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.023$, $\Delta_{A_1B_1} = 0.06$ and $\Delta_{A_1D_1B_1} = 0.013$
 average of 20,000 replicates each with N cases and N controls, 0.05 significance level; $p_{D_1} = 0.15$, $\phi = 0.1$ and $\lambda_2 = 1.4$; S_{LD}^m and S_{LD}^w are calculated under the assumption that the haplotype frequencies in cases and controls are the same. S_{LD}^m and S_{LD}^w are almost equal, though S_{LD}^m is slightly better than S_{LD}^w .

Figure 5. Empirical power when $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.047$, $\Delta_{A_1B_1} = 0.06$ and $\Delta_{A_1D_1B_1} = 0.03$
 average of 20,000 replicates each with N cases and N controls, 0.05 significance level; $p_{D_1} = 0.15$, $\phi = 0.1$ and $\lambda_2 = 1.4$; S_{LD}^m and S_{LD}^w are calculated under the assumption that the haplotype frequencies in cases and controls are the same. In panels **a**, **b** and **d**, S_{LD}^m and S_{LD}^w are almost equal, but S_{LD}^m is slightly better than S_{LD}^w in panels **a** and **b**, and vice versa in panel **d**.

Figure 6. Empirical power when $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.068$, $\Delta_{A_1B_1} = 0.06$ and $\Delta_{A_1D_1B_1} = 0.049$
 average of 20,000 replicates each with N cases and N controls, 0.05 significance level; $p_{D_1} = 0.15$, $\phi = 0.1$ and $\lambda_2 = 1.4$; S^m_{LD} and S^w_{LD} are calculated under the assumption that the haplotype frequencies in cases and controls are the same. In panels **a** and **b**, S^m_{LD} and S^w_{LD} are almost equal, but S^m_{LD} is slightly better than S^w_{LD} , and S^c_{LD} and S^w_{LD} are almost equal in panel **d**.

Figure 7. Correlations

p_{A1} and p_{B1} are 0.2; D' , Lewontin's D'^5 ; **a** and **b** are analytical correlations of p_{A1} and d_A with $\widehat{\Delta}_{A1B1}^m$, $\widehat{\Delta}_{A1B1}^w$ and $\widehat{\Delta}_{A1B1}^c$ when the haplotype frequencies are known; **c** and **d** are empirical correlations of S_a and S_{HWD} with S_{LD}^m , S_{LD}^w and S_{LD}^c from 20,000 replicates of 10,000 cases and 10,000 controls under LE between marker and disease alleles; $p_{A1} = 0.15$, $\phi = 0.1$, $\lambda_1 = 1.2$ and $\lambda_2 = 1.4$; S_{LD}^m and S_{LD}^w are calculated under the assumption that the haplotype frequencies in cases and controls are the same.

Figure 8. SED as a function of ΔA_{1D}

$p_{A1} = p_{B1} = 0.2$ and $p_{D1} = 0.1$, homozygote disease genotype relative risk is 1.4, and the prevalence of the disease is 0.1; the heterozygote disease genotype relative risk is calculated according to the disease mode of inheritance and the marker and disease allele frequencies and S_{LD}^m and S_{LD}^w are calculated under the assumption that the haplotype frequencies in cases and controls are the same. S_{LD} , S_{LD}^m and S_{LD}^w are almost equal, though S_{LD} is the best, followed by S_{LD}^m , in the panels **a**, **b**, and **c**.

Figure 9. SED as a function of $\Delta_{A_1B_1}$

$p_{A_1} = p_{B_1} = 0.2$ and $p_{D_1} = 0.1$, homozygote disease genotype relative risk is 1.4, and the prevalence of the disease is 0.1; $\Delta_{A_1D_1} = 0.076$, $\Delta_{D_1B_1} = 0.068$, $\Delta_{A_1B_1} = 0.06$ and $\Delta_{A_1D_1B_1} = 0.049$; the heterozygote disease genotype relative risk is calculated according to the disease mode of inheritance and S^m_{LD} and S^w_{LD} are calculated under the assumption that the haplotype frequencies in cases and controls are the same.

Figure 10. SED as a function of λ_1

λ_2 is assumed to be 3, and both the prevalence and disease allele frequency are 0.1; $\Delta_{A_1D_1} = 0.09$ ($D'=1$), $\Delta_{A_1D_1B_1} = 0.1\Delta_{A_1D_1}$ and $\Delta_{B_1D_1} = 0.2\Delta_{A_1D_1}$; the two vertical lines indicate λ_1 for additive and multiplicative disease mode of inheritance.

Simulations with sparse marker data

Table 1

the empirical type I error and the empirical power (2,000 cases and 2,000 controls) for each disease mode of inheritance at the nominal 0.05 and 0.01 significance levels, for sparse marker data from Hapmap; for S_{LD} , the most probable haplotypes are assumed to be the true haplotypes.

Statistic	S_a		S_{HWD}		S_{LD}		S_{LD}^C	
Significance level	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
Empirical type I error	0.047	0.017	0.05	0.008	0.016	0.001	0.057	0.008
add	0.415	0.196	0.056	0.013	0.039	0.008	0.053	0.016
mul	0.35	0.148	0.056	0.014	0.032	0.004	0.061	0.012
dom	0.81	0.603	0.065	0.016	0.123	0.027	0.082	0.022
rec	0.084	0.016	0.05	0.006	0.025	0.001	0.051	0.008

Simulations with dense marker data

Table 2

the empirical type I error and the empirical power (2,000 cases and 2,000 controls) for each disease mode of inheritance at the nominal 0.05 and 0.01 significance levels, for dense marker data from Hapmap; for S_{LD} , the most probable haplotypes are assumed to be the true haplotypes.

Statistic	S_a		S_{HWD}		S_{LD}		S_{LD}^C	
Significance level	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
Empirical type I error								
	0.045	0.009	0.05	0.014	0.044	0.01	0.053	0.014
	Add	0.725	0.495	0.06	0.013	0.869	0.701	0.426
	Mul	0.653	0.413	0.048	0.011	0.825	0.626	0.397
	Dom	0.97	0.903	0.12	0.035	0.999	0.986	0.675
Power	Rec	0.178	0.065	0.1	0.039	0.233	0.094	0.219

Table 3
Effect of maximum possible HWD in the population under LE

empirical type I error at the nominal 0.05 and 0.01 significance levels, with the variances of the statistics calculated assuming HWE or estimating HWD separately in cases and controls; for each disease mode of inheritance we assume $\Delta_{A_1D} = 0$.

Statistic	S_a				S_{HWD}				S_{LD}^C			
Significance level	0.05	0.01	0.01	0.05	0.05	0.01	0.01	0.05	0.05	0.01	0.01	0.05
HWD	Dom	0.0480	0.011	0.051	0.051	0.011	0.011	0.051	0.051	0.010	0.010	
	Rec	0.050	0.010	0.053	0.053	0.012	0.012	0.050	0.050	0.011	0.011	
	Add	0.049	0.009	0.052	0.052	0.010	0.010	0.049	0.049	0.010	0.010	
	Mul	0.047	0.010	0.050	0.050	0.010	0.010	0.050	0.050	0.010	0.010	
HWE	Dom	0.056	0.011	0.052	0.052	0.011	0.011	0.057	0.057	0.012	0.012	
	Rec	0.049	0.010	0.055	0.055	0.012	0.012	0.057	0.057	0.012	0.012	
	Add	0.048	0.010	0.054	0.054	0.012	0.012	0.057	0.057	0.012	0.012	
	Mul	0.050	0.009	0.054	0.054	0.012	0.012	0.053	0.053	0.011	0.011	